Mapping Short Reads to a Genomic Sequence with Circular Structure

Tomas Flouri, Czech Technical University in Prague, Czech Republic
Costas S. Iliopoulos, King’s College London, UK, and Curtin University, Australia
Solon P. Pissis, King’s College London, UK
German Tischler, University of Wuerzburg, Germany

ABSTRACT

Constant advances in DNA sequencing technologies are turning whole-genome sequencing into a routine procedure, resulting in massive amounts of data that need to be processed. Tens of gigabytes of data, in the form of short sequences (reads), need to be mapped back onto reference sequences, a few gigabases long. A first generation of short-read alignment algorithms successfully employed hash tables, and the current second generation uses the Burrows-Wheeler transform, further improving speed and memory footprint. These next-generation sequencing technologies allow researchers to characterise a bacterial genome, during a single experiment, at a moderate cost. In this article, as most of the bacterial chromosomes contain a circular DNA molecule, the authors present a new simple, yet efficient, sensitive and accurate algorithm, specifically designed for mapping millions of short reads to a genomic sequence with circular structure.

Keywords: Algorithms on Strings, Circular Pattern Matching, DNA Technologies, Genomic Sequence, Next-Generation Sequencing

INTRODUCTION

Sequencing technology has come a long way since the time when traditional sequencing techniques required many labs around the world to cooperate for years in order to sequence the human genome for the first time. The traditional Sanger-based sequencing methods, developed in the mid 70’s, had been the workhorse technology for DNA sequencing for almost 30 years (Sanger & Coulson, 1975; Sanger et al., 1977). Nowadays, next-generation sequencing technologies have reduced the task of sequencing a whole genome to a matter of days, or even hours, and the cost has decreased by orders of magnitude, making it an accessible experimental procedure to many labs (ten Bosch & Grody, 2008). This opened the door for re-sequencing to start becoming a more routine procedure, as it finds many applications in the detection of genetic variability among individuals. Thus,
it can help us understand the extent of that variability, and also identify specific variants, alternative splicing sites and patterns, epigenetic effects, and relate them to gene regulation and expression, as well as to diseases (1000 Genomes, 2011; Wu & Nacu, 2010; Xiang et al., 2010; Ng et al., 2010). Thus, DNA sequencing is quickly becoming a powerful tool in diagnostic medicine, and eventually personalised treatment (ten Bosch & Grody, 2008).

The data resulting from a single sequencing experiment can be massive; it is not uncommon to have data from multiple experiments. This trend of increasing availability of sequencing data will continue as projects even more ambitious than the 1000 Genomes Project (1000 Genomes, 2011) start to materialize. According to their respective websites, typical output sizes of the three main next-generation sequencing platforms – 454/Roche, ABI SOLiD, and Illumina GA – are millions of reads ranging in size from 25bp to 400bp. In most cases, these reads are too short to be directly assembled, especially in the presence of repetitive regions (Miller et al., 2010), therefore a reference sequence is usually required.

Mapping so many short reads onto a reference sequence is a very challenging task that cannot be adequately carried out by traditional search and alignment algorithms (Kent, 2002) like BLAST (Altschul et al., 1990) and FASTA (Pearson & Lipman, 1988), so a broad array of programmes (Jiang & Wong, 2008; Li et al., 2009; Langmead et al., 2009; Li & Durbin, 2009; Frousios et al., 2010) has been published to address this task, placing emphasis on different aspects of the challenge. The different algorithms implement various combinations of innovations and trade-offs, to address computing speed, system resources requirements, and biological relevance and accuracy of the computed results.

Unlike the linear DNA of vertebrates, strain or species of bacteria with circular organization of their chromosomes or plasmids, are the most common. Until towards the end of the 1980s, when the technology for examining chromosomes and plasmids improved, all bacteria were thought to have a single circular chromosome (Colem & Saint-Girons, 1999). In fact, not all bacteria have a single circular chromosome; some bacteria have multiple circular chromosomes (Suwanto & Kaplan, 1989a, 1989b, 1992a, 1992b), and many bacteria have linear chromosomes and linear plasmids (Volf & Altenbuchner, 2000). Bacterial genomes range in size from about 160, 000bp to 12, 200, 000bp, depending on the type considered (Nakabachi et al., 2006).

In this article, as most of the bacterial chromosomes contain a circular DNA molecule, we present a new simple, yet efficient, sensitive and accurate algorithm, specifically designed for mapping millions of short reads onto a genomic sequence with circular structure. In particular, we define this problem as a circular pattern matching (CPM) problem, reduce it to a classical pattern matching problem, and make use of REAL, the read aligner proposed in Frousios et al. (2010), to efficiently map the reads to the genomic sequence.

The remainder of the paper is structured as follows. Section Basic Definitions presents the basic definitions that are used throughout the paper. In Section Problem Definition, we formally define the problem solved in this paper. Section The CPM algorithm presents the CPM algorithm for mapping short reads onto a genomic sequence with circular structure. Finally, in Section Experimental Results, we present extensive experimental results, which demonstrate the importance of the proposed approach compared to more traditional approaches, and we briefly conclude in Section Conclusion.

**BASIC DEFINITIONS**

A string or sequence is a succession of zero or more symbols from an alphabet \( \Sigma \); the string with zero symbols is denoted by \( \epsilon \). The set of all strings over the alphabet \( \Sigma \), including \( \epsilon \), is denoted by \( \Sigma^* \). The set \( \Sigma^+ \) is defined as \( \Sigma^* \setminus \{ \epsilon \} \). A string \( x \) of length \( m \) is represented by \( x[0..m-1] \), where \( x[i] \in \Sigma \).
Related Content

Aβ Monomer, Oligomer and Fibril in Alzheimer’s Disease
www.igi-global.com/chapter/monomer-oligomer-fibril-alzheimer-disease/76133?camid=4v1a

Privacy Preserving Principal Component Analysis Clustering for Distributed Heterogeneous Gene Expression Datasets
Xin Li (2013). Methods, Models, and Computation for Medical Informatics (pp. 238-271).
www.igi-global.com/chapter/privacy-preserving-principal-component-analysis/73081?camid=4v1a

The Applications of Omics Technologies and the Challenges of Ethics in Nutritional Sciences
www.igi-global.com/chapter/applications-omics-technologies-challenges-ethics/47294?camid=4v1a
Predicting Protein Functions from Protein Interaction Networks
Hon Nian Chua and Limsoon Wong (2012). International Journal of Knowledge Discovery in Bioinformatics (pp. 50-70).
www.igi-global.com/article/predicting-protein-functions-from-protein-interaction-networks/101242?camid=4v1a